Теорія Ймовір. та матем. статист.
Вип. 72, 2005, стор. 111–119

Teor. Ĭmovir. ta Matem. Statyst.
No. 72, 2005, pp 111–119

ГІСТОГРАМНІ ОЦІНКИ ФОРМИ ФУНКЦІЇ КОНЦЕНТРАЦІЇ ДВОКОМПОНЕНТНОЇ СУМІШІ

УДК 519.21

Д. І. ПОХИЛЬКО

Анотація. В роботі розглянуто питання про побудову проекційних оцінок форми функції концентрації з розкладом за базисом індикаторних функцій (гістограмним базисом) для даних, що являють собою вибірку з суміші двох компонент з невідомими розподілами, концентрація яких змінюється протягом спостережень. Доведено консистентність оцінок і оцінено швидкість збіжності м.н.

1. Вступ

Суміші різних ймовірнісних розподілів часто використовуються у статистиці як математичні моделі розподілу реальних даних. Класична модель суміші має вигляд

$$\mathsf{P}\{\xi \in A\} = w^1 H_1(A) + w^2 H_2(A) + \cdots + w^M H_M(A)$$

де M — кількість компонент суміші, H_m — деякі ймовірнісні розподіли, w^m — концентрація m-тої компоненти у суміші. Як правило, їх використовують, коли об'єкти, що спостерігаються, можуть належати одному з M різних класів (популяцій, компонент), розподіли їх спостережуваних характеристик (ξ) дорівнюють H_m для m-тої компоненти, а ймовірність спостерігати об'єкт з m-тої компоненти дорівнює w^m. Якщо невідомо, до якої популяції належать спостережувані об'єкти, ми отримуємо в точності класичну модель суміші, (див. [5] та список літератури в цій роботі).

Питання про побудову оцінок розподілу при відомих концентраціях розглядається в [1, 4]. Задачі оцінки щільності розподілу і класифікації по сумішах зі змінними концентраціями розглядаються в [10, 11].

Ми розглянемо задачу оцінювання функції концентрації у випадку, коли концентрації компонент змінюються від спостереження до спостереження, і суміш складена з двох компонент, розподіли яких невідомі. Тобто, спостереження ξ_j є незалежними в сукупності і

$$\mathsf{P}\{\xi_j \in A\} = w(t_j)H_1(A) + (1 - w(t_j))H_2(A),$$

де $w(t_j)$ — концентрація 1-ої компоненти суміші в момент часу t_j коли проводилось j-те спостереження. Цю задачу можна було б розв'язувати з використанням методів непараметричної регресії (див. в [8]). Дійсно, якщо η_1, η_2 — в.в. з скінченим математичним сподіванням, що описують характеристики компонент, то ξ_j допускає зображення $\xi_j = w(t_j)(\mathsf{E}\,\eta_1 - \mathsf{E}\,\eta_2) + \mathsf{E}\,\eta_2 + \varepsilon_j$, де ε_j — "шум" (незалежні центровані в.в.). Але при такому підході ми втрачаємо важливу інформацію про $w(t_j)$, яка міститься в ε_j. Наприклад, якщо $\mathsf{E}\,\eta_1 = \mathsf{E}\,\eta_2$, то на основі регресійного підходу не можна оцінити $w(t_j)$. Оцінка, розглянута в даній роботі, є консистентною і при $\mathsf{E}\,\eta_1 = \mathsf{E}\,\eta_2$, для такого випадку побудовано приклад оцінки в п. 3.

2000 *Mathematics Subject Classification.* Primary 62G20; Secondary 62G05.
Ключові слова і фрази. Проекційні оцінки, оцінки концентрації, форма функції концентрації.

Проектування на гістограмний базис є частковим випадком застосування техніки проекції на вейвлет-базиси (див. [7, 6]). Використання вейвлетів в статистиці розглянуто в [9, 12]. В [9] розглянуто питання про застосування вейвлетів до задач непараметричної регресії.

Оцінки, розглянуті в роботі, аналогічні проекційним оцінкам, розглянутим в [2, 3], побудованим з розкладом по тригонометричному базису. В [3] побудовано оцінки для функцій розподілу компонент суміші, які можна отримати, маючи оцінку функції концентрації.

2. Постановка задачі і побудова оцінки

Нехай $\{\xi_j^N, j = 1, \dots, N\}_{N\geq 1}$ — послідовність серій випадкових векторів в $\mathbb{R}^d$, таких, що для всіх N вектори $\{\xi_j^N, j = 1, \dots, N\}$ — незалежні в сукупності,

$$\mathsf{P}\{\xi_j^N < x\} = F(t_j, x) = w(t_j)G(x) + (1 - w(t_j))H(x), \tag{1}$$

де G, H — функції розподілу (ф.р.) на $\mathbb{R}^d$, $w\colon [0,1] \to [0,1]$ — концентрація компоненти G в час $t_j = j/N$. Позначимо Lip_β — клас Ліпшиця порядку β, тобто $\mathrm{Lip}_\beta[0,1] = \{f\colon [0,1] \to \mathbb{R} \mid \text{існує } L'\colon \sup_{t,s\in[0,1]} |f(t) - f(s)| \leq L'|t-s|^\beta\}$,

$$\Sigma(\beta, L') = \left\{ f\colon [0,1] \to \mathbb{R} \mid \sup_{t,s\in[0,1]} |f(t) - f(s)| \leq L'|t-s|^\beta \right\}.$$

В цій роботі розглянуто задачу оцінювання форми w за спостереженнями ξ_j^N в припущенні $w \in \mathrm{Lip}_\beta[0,1]$, де $\beta \in [0,1]$ — фіксоване (відоме), G, H — невідомі.

Функцію $F(t,x)$ можна подати у вигляді

$$F(t,x) = G_1(x) + u(t)G_2(x), \tag{2}$$

де

$$G_1(x) = H(x) + abG_2(x), \qquad G_2(x) = \frac{G(x) - H(x)}{a},$$

$$u(t) = a(w(t) - b), \qquad b = \int_0^1 w(t)\,dt, \qquad a = \left(\int_0^1 (w(t) - b)^2\,dt\right)^{-1/2}.$$

Зрозуміло, що

$$\int_0^1 (u(t))^2\,dt = 1, \qquad \int_0^1 u(t)\,dt = 0. \tag{3}$$

Якщо $w(t)$ — функція концентрації, то $u(t)$ — її форма. Як показано в [2], розклад (2) при виконанні умов нормування (3) стає однозначним з точністю до знаку $u(t)$. Якщо не вимагати виконання (3), то розклад стає неоднозначним. Тому надалі будемо оцінювати $u(t)$, визначене (2)–(3). Таким чином, ми оцінюємо форму функції концентрації, тобто функцію концентрації з точністю до зсуву та масштабу.

Ми будемо шукати функцію $u_N^*(t)$, яка добре наближує u на рівномірній сітці часу, тому на роль критерія якості підгонки оберемо

$$\sqrt{\frac{1}{N}\sum_{k=1}^N \left(u\left(\frac{k}{N}\right) - u_N^*\left(\frac{k}{N}\right)\right)^2}.$$

При побудові оцінки $u_N^*(t)$ використаємо підхід, запропонований в [2]. Для цього будемо проектувати $F(t,x)$ на клас функцій

$$F_v(t,x) = G_1(x) + v(t)G_2(x),$$

де $v \in V$, V — скінченовимірний клас функцій в $L_2([0,1], m)$, (m тут і надалі — міра Лебега), причому v задовольняє умовам нормування (3). Нехай $\pi(\cdot)$ — деяка ймовірнісна міра на $\mathbb{R}^d$, Δ — носій міри π. Шукатимемо v, яке задовольняє (3) і мінімізує величину

$$\sqrt{\frac{1}{N}\int_\Delta \sum_{k=1}^N \left(F\left(\frac{k}{N},x\right)-F_v\left(\frac{k}{N},x\right)\right)^2 \pi(dx)}. \tag{4}$$

Позначимо λ_N — ймовірностну рахуючу міру, що зосереджена в точках $t_j^N = j/N$, $j=1,\dots,N$, тобто $\lambda_N(\{t_j^N\}) = 1/N$. Тут і надалі під $\|\cdot\|_N$, $\langle\cdot,\cdot\rangle_N$ будемо розуміти, відповідно, норму і скалярний добуток в просторі $L_2([0,1],\lambda_N)$, під $\|\cdot\|_2$ — норму в $L_2([0,1],m)$, тоді в цих позначеннях

$$\sqrt{\frac{1}{N}\sum_{k=1}^N \left(u(t_k^N)-u_N^*(t_k^N)\right)^2} = \|u-u_N^*\|_N.$$

Позначимо $\mathbb{H} = L_2([0,1]\times\Delta, \lambda_N\times\pi)$, тоді задача мінімізації (4), тобто мінімізації $\|F_v - F\|_{\mathbb{H}}$ на V еквівалентна максимізації

$$\begin{aligned} D(v) &= \int_\Delta \left(\int_{[0,1]} v(t)F(t,x)\,\lambda_N(dt)\right)^2 \pi(dx) \\ &= \frac{1}{N^2}\int_\Delta\int_\Delta \left(\sum_{k=1}^N v\left(\frac{k}{N}\right)F\left(\frac{k}{N},x\right)\right)^2 \pi(dx). \end{aligned} \tag{5}$$

Оскільки функція розподілу $F(t,x)$ нам невідома, то шукатимемо оцінку для v, підставляючи в (5) оцінку для F. Зрозуміло, що побудова оцінки буде залежати лише від вибору V і, наприклад, якщо $\{f_i, i=1,\dots,l\}$ — деякий ортонормований базис (ОНБ) з.л.о.(V) в $L_2([0,1],\lambda_N)$, то оцінку u_N^* можна побудувати так

$$u_N^* = \sum_{i=1}^l c_N^i f_i, \tag{6}$$

де $c_N = (c_N^i, i=1,\dots,l)$ — власний вектор матриці $(d_{ij}^N)_{i,j=1}^l$, що відповідає найбільшому власному числу, де

$$d_{ij}^N = \frac{1}{N^2}\sum_{l,k=1}^N f_i(t_l^N)f_j(t_k^N)\int_\Delta \mathbb{I}\{\xi_l^N < x, \xi_k^N < x\}\,\pi(dx), \tag{7}$$

(див. [2, 3]).

Введемо наступні позначення:

$$g_k^l(t) := \sqrt{l+1}\,\mathbb{I}\left\{t\in\left(\frac{k}{l+1},\frac{k+1}{l+1}\right]\right\}, \qquad V_l = \text{з.л.о.}\{g_k^l, k=0,\dots,l\},$$

замикання береться в просторі $L_2((0,1],\lambda_N)$; $U_l = \text{з.л.о.}\{g_k^l, k=0,\dots,l\}$, замикання береться в просторі $L_2((0,1],m)$, $\{g_k^l, k=0,\dots,l\}$ — ОНБ в U_l. Нехай

$$\{f_j^{Nl}(t), j=0,\dots,l\}$$

— ОНБ в V_l відносно скалярного добутку $\langle\cdot,\cdot\rangle_N$, при цьому покладемо $f_0^{Nl}\equiv 1$, $W_l = \text{з.л.о.}\{f_k^{Nl}, k=1,\dots,l\}$, замикання береться в просторі $L_2((0,1],\lambda_N)$.

Теорема 2.1. *Нехай виконано (2) з обмеженнями (3),*

$$\int_\Delta (G_2(x))^2\,\pi(dx) > 0, \qquad u\in\Sigma(\beta,L'),\ \beta\in(0,1],$$

оцінка u_N^* визначена (6), (7), де в якості V виступає W_l. Якщо $l = N^\alpha$, $\alpha\in(0,\frac{1}{2})$, то

(1) $\min_{\sigma=\pm1}\|u-\sigma u_N^*\|_N = O(r_N(\alpha,\beta))$ *м.н.*, $N\to\infty$, *де*

$$r_N(\alpha,\beta) = (\ln N)^{1/4}N^{\alpha/2-1/4} + N^{-\alpha\beta};$$

(2) $\min_{\sigma=\pm1}\|u-\sigma u_N^*\|_2 = O(r_N(\alpha,\beta))$ *м.н.*, $N\to\infty$.

Нам будуть потрібні наступні леми.

Лема 2.1. *Нехай* $z=(z^1,\dots,z^l)\in\mathbb{R}^l$, $\|z\|\in[\frac{1}{2},2]$, $u_z=\sum_{i=1}^l z^i f_i^{Nl}$,

$$F(t,x) = G_1(x) + u_z(t)G_2(x) + \Delta_N(t,x),$$

де $\Delta_N(t,x)$ *— довільна функція. Для довільного* $\zeta\in\mathbb{R}^l$, *такого що* $\|\zeta\|=1$ *(норма в* $\mathbb{R}^l$*) покладемо*

$$\Delta_\zeta(x) = \int_{[0,1]} u_\zeta(t)\Delta_N(t,x)\,\lambda_N(dt), \qquad C = \int_\Delta (G_2(x))^2\,\pi(dx),$$

$$D(\zeta) := D(u_\zeta) = \int_\Delta\left(\int_{[0,1]} u_\zeta(t)(G_1(x)+u_z(t)G_2(x)+\Delta_N(t,x))\,\lambda_N(dt)\right)^2\pi(dx),$$

$$R(\zeta) = \int_\Delta G_2(x)\Delta_\zeta(x)\,\pi(dx), \qquad Q(\zeta) = \int_\Delta (\Delta_\zeta(x))^2\,\pi(dx),$$

$$\Delta_\zeta = 4|R(\zeta)| + Q(\zeta), \qquad \delta=\delta_N=\sup_{\|\zeta\|=1}\Delta_\zeta, \qquad v=\operatorname{argmax}_{\|\zeta\|=1} D(\zeta),$$

тоді

$$\min_{\sigma=\pm1}\|u_z-\sigma u_v\|_N \le \big|\|z\|-1\big| + 4\sqrt{\delta/C}.$$

Зауважимо, що $\|z\|=\|u_z\|_N$. Дана лема є модифікацією леми 1 в [3].

Доведення. Можна показати, що

$$D(\zeta) = C\,\langle u_\zeta,u_z\rangle_N^2 + 2\,\langle u_\zeta,u_z\rangle_N\,R(\zeta) + Q(\zeta).$$

Покладемо $y=z/\|z\|$, тоді $\|y\|=1$, $z=y\|z\|$. Отримаємо подання

$$D(\zeta) = C\,\|z\|^2\,\langle u_\zeta,u_y\rangle_N^2 + 2R(\zeta)\,\|z\|\,\langle u_\zeta,u_y\rangle_N + Q(\zeta).$$

Аналогічно тому, як це зроблено в [3], можна отримати

$$\min_{\sigma=\pm1}\|u_y-\sigma u_v\|_N \le 4\sqrt{\frac{\delta}{C}}.$$

Тому з нерівності трикутника отримаємо

$$\min_{\sigma=\pm1}\|u_z-\sigma u_v\|_N \le \|u_z-u_y\|_N + \min_{\sigma=\pm1}\|u_y-\sigma u_v\|_N \le \big|\|z\|-1\big| + 4\sqrt{\frac{\delta}{C}}. \quad \square$$

Наступна лема доводиться в [2, Л.3.6.3.].

Лема 2.2. *Якщо виконуються умови теореми 2.1,*

$$u(t) = \sum c_{i=1}^l f_i^{Nl}(t) + u^0(t),$$

де $\|c\|=1$ *(норма вектора в* $\mathbb{R}^l$*),* $\langle u^0, f_i^{Nl}\rangle_N = 0$, $i=1,\dots,l$,

$$\Delta_N(t,x) = \mathbb{1}\{\xi_{k(t)}^N < x\} - G_1(x) - \sum_{i=1}^N c^i f_i^{Nl}(t)G_2(x), \tag{8}$$

$k(t)=\min\{y\in\mathbb{Z}\mid y/N\ge t\}$, *то в позначеннях леми 2.1*

$$\delta_N = O\big((\ln N)^{1/2}\big)N^{\alpha-0.5} \quad \text{м.н.}$$

Зауважимо, що при такому, як в (8), виборі $\Delta_N(t,x)$ функція u_v з леми 2.1 переходить в оцінку u_N^*, визначену в (6).

Застосувавши теорему про середнє для неперервних функцій, можна отримати наступне твердження:

Лема 2.3. *Нехай* $u\colon [0,1] \to \mathbb{R}$, $u \in \Sigma(\beta, L')$. *Тоді* $\sup_{t\in(0,1]} |u(t) - v(t)| \le L' l^{-\beta}$, *де* $v(t) = \Pr_{U_l} u(t)$, *проекція будується в просторі* $L_2((0,1], m)$.

Лема 2.4. *Нехай* $N \ge 4$, $l = N^\alpha$, $\alpha \in (0, \frac{1}{2})$; $u, v\colon [0,1] \to \mathbb{R}$; $u \in \Sigma(\beta, L')$,

$$v(t) = \sum_{k=0}^{l} c_k \mathbb{I}\left\{t \in \left(\frac{k}{l+1}, \frac{k+1}{l+1}\right]\right\}.$$

Тоді існують $K_i = K_i(L') \in \mathbb{R}$, $i = 1,2,3$:

$$\|u - v\|_2 \le K_1 N^{-\beta} + K_2 \|u - v\|_N + K_3 N^{-\beta/2} \|u - v\|_N^{1/2}.$$

Доведення. Маємо

$$\begin{aligned}
\|u - v\|_2^2 &= \int_0^1 (u(t) - v(t))^2\,dt = \sum_{k=0}^{l} \int_{\frac{k}{l+1}}^{\frac{k+1}{l+1}} (u(t) - v(t))^2\,dt \\
&\le \sum_{k=0}^{l} \sum_{s\in\mathbb{Z}:\, (\frac{s}{N}, \frac{s+1}{N}] \subset (\frac{k}{l+1}, \frac{k+1}{l+1}]} \int_{\frac{s}{N}}^{\frac{s+1}{N}} (u(t) - v(t))^2\,dt \\
&\quad + \sum_{k=0}^{l} \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} (u(t) - v(t))^2\,dt \\
&= S_1 + S_2,
\end{aligned}$$

де s_k обираються з умови $s_k \in \mathbb{Z}$,

$$\frac{Nk - 1}{l+1} < s_k \le \frac{kN}{l+1}.$$

При такому підсумовуванні в першій сумі (в S_1) для всіх $t \in (s/N, (s+1)/N]$

$$v(t) = v\left(\frac{s}{N}\right),$$

а всі точки, де $v(t)$ може змінити значення всередині проміжку інтегрування, віднесені в другу суму (S_2). Вимагатимемо $N \ge 4$, щоб всередині інтервалу довжиною $1/N$ функція $v(t)$ не змінювала своє значення більше одного разу.

Оцінимо S_1. Враховуючи, що для доданків з S_1 для всіх $t \in (s/N, (s+1)/N]$,

$$v(t) = v\left(\frac{s+1}{N}\right),$$

можемо мажорувати їх величиною

$$\frac{1}{N}\left(u\left(\frac{s+1}{N}\right) - v\left(\frac{s+1}{N}\right)\right)^2 + \frac{2L'N^{-\beta}}{N}\left|u\left(\frac{s+1}{N}\right) - v\left(\frac{s+1}{N}\right)\right| + \frac{L'^2 N^{-2\beta}}{N}.$$

Для S_2

$$\begin{aligned}
&\int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} (u(t) - v(t))^2\,dt \\
&\quad \le \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} \left(u(t) - v\left(\frac{s_k}{N}\right)\right)^2 dt + \int_{\frac{s_k}{N}}^{\frac{s_k+1}{N}} \left(u(t) - v\left(\frac{s_k+1}{N}\right)\right)^2 dt = I_1 + I_2.
\end{aligned}$$

Величину I_2 вже оцінено вище, оцінимо I_1. Використавши теорему про середнє для u, можна показати, що

$$I_1 \le \frac{L'^2 N^{-2\beta}}{N} + \frac{2L' N^{-\beta}}{N} \int_{\frac{s_{k-1}}{N}}^{\frac{s_k}{N}} \left| u(t) - v\left(\frac{s_k}{N}\right) \right| dt + \int_{\frac{s_{k-1}}{N}}^{\frac{s_k}{N}} \left(u(t) - v\left(\frac{s_k}{N}\right) \right)^2 dt.$$

Отже, існують такі $C_1, C_2, C_3 \in \mathbb{R}$, що

$$\begin{aligned} S_1 + S_2 &\le C_1 N^{-2\beta} + \frac{C_2}{N} \sum_{k=1}^{N} \left(u\left(\frac{k}{N}\right) - v\left(\frac{k}{N}\right) \right)^2 + \frac{C_3 N^{-\beta}}{N} \sum_{k=1}^{N} \left| u\left(\frac{k}{N}\right) - v\left(\frac{k}{N}\right) \right| \\ &\le C_1 N^{-2\beta} + C_2 \|u - v\|_N^2 + C_3 N^{-\beta} \|u - v\|_N , \end{aligned}$$

що і доводить лему. □

Доведення теореми 2.1. Скористаємося нерівністю трикутника:

$$\min_{\sigma=\pm 1} \|u - \sigma u_l^*\|_N \le \|u - \hat{u}_l\|_N + \|\hat{u}_l - \tilde{u}_l\|_N + \min_{\sigma=\pm 1} \|\tilde{u}_l - \sigma u_N^*\|_N ,$$

де $\hat{u}_l = \mathrm{Pr}_{V_l} u$, $\tilde{u}_l = \mathrm{Pr}_{W_l} u$, проекція будується в просторі $L_2((0,1], \lambda_N)$. Оцінимо перший доданок. Використавши позначення леми 2.3 і те, що $v_l \in V_l$, де $v_l = \mathrm{Pr}_{U_l} \hat{u}_l$, можемо записати:

$$\|u - \hat{u}_l\|_N = \min_{y \in V_l} \|u - y\|_N \le \|u - v_l\|_N \le \sup_{t \in (0,1]} |u(t) - v_l(t)| \le L' N^{-\alpha\beta}.$$

Оцінимо другий доданок. Оскільки $W_l \subset V_l$ і $\int_{[0,1]} u(t)\, dt = 0$, то

$$\begin{aligned} \|\hat{u}_l - \tilde{u}_l\|_N &= \|\mathrm{Pr}_{V_l \ominus W_l} u\|_N = |\langle u, f_0^{Nl} \rangle_N| = \left| \int_{[0,1]} u(t)\, \lambda_N(dt) \right| \\ &= \left| \int_{[0,1]} u(t)\, dt - \int_{[0,1]} u(t)\, \lambda_N(dt) \right| \le \sum_{k=1}^{N} \int_{\frac{k-1}{N}}^{\frac{k}{N}} \left| u(t) - u\left(\frac{k}{N}\right) \right| dt \le L' N^{-\beta}. \end{aligned}$$

Оцінимо величину $|1 - \|\tilde{u}_l\|_N|$:

$$\begin{aligned} |1 - \|\tilde{u}_l\|_N| &\le |1 - \|u\|_N| + |\|u\|_N - \|\hat{u}_l\|_N| + |\|\hat{u}_l\|_N - \|\tilde{u}_l\|_N| \\ &\le |1 - \|u\|_N| + \|u - \hat{u}_l\|_N + \|\hat{u}_l - \tilde{u}_l\|_N \le |1 - \|u\|_N| + L' N^{-\alpha\beta} + L' N^{-\beta}, \end{aligned}$$

$$\begin{aligned} |1 - \|u\|_N| = \frac{|1 - \|u\|_N^2|}{1 + \|u\|_N} &\le |\|u\|_2^2 - \|u\|_N^2| = \left| \sum_{k=1}^{N} \int_{\frac{k-1}{N}}^{k/N} \left(u^2(t) - u^2\left(\frac{k}{N}\right) \right) dt \right| \\ &\le 2 \sup_{t \in (0,1]} |u(t)| \frac{1}{N} \sum_{k=1}^{N} \sup_{t \in \left(\frac{k-1}{N}, \frac{k}{N}\right]} \left| u(t) - u\left(\frac{k}{N}\right) \right| \le 2 L'^2 N^{-\beta}. \end{aligned}$$

Тут було використано те, що для нормованих просторів $|\|a\| - \|b\|| \le \|a - b\|$ і те, що з $u \in \Sigma(\beta, L')$ і

$$\int_{[0,1]} u(t)\, dt = 0$$

випливає $\sup_{t \in [0,1]} |u(t)| \le L'$. Таким чином

$$|1 - \|\tilde{u}\|_N| \le C_1' N^{-\beta} + L' N^{-\alpha\beta},$$

тому при достатньо великих N виконуються умови леми 2.1, і отже:

$$\min_{\sigma=\pm 1} \|\tilde{u}_l - \sigma u_N^*\|_N \le C_1' N^{-\beta} + L' N^{-\alpha\beta} + 4\sqrt{\frac{\delta_N}{C}},$$

$$\min_{\sigma=\pm 1} \|u - \sigma u_N^*\|_N \le C_1 N^{-\beta} + C_2 N^{-\alpha\beta} + C_3 \sqrt{\delta_N},$$

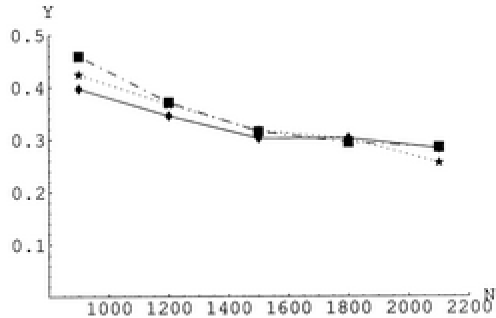

Рис. 1. Усереднена відстань $\min_{\sigma=\pm1}\|u-\sigma u_N^*\|_2$ для прикладу A

де δ_N оцінена в лемі 2.2, що і доводить перше твердження теореми, друге випливає з першого і леми 2.4. □

Швидкість збіжності, отримана в цій теоремі, співпадає з швидкістю збіжності для оцінок, побудованих з розкладом по тригонометричному базису (див. [3]).

3. Поведінка оцінки на модельованих даних

Розглянемо два приклади. Нехай для обох $w(t)=4(t-\frac{1}{2})^2\mathbb{1}\{t\in[0,1]\}$, для прикладу A) $\pi(A)=m(A\cap[0,1])$ для всіх $A\in\mathcal{B}(\mathbb{R})$; для прикладу B) $\pi(A)=\frac{1}{2}m(A\cap[0,2])$ для всіх $A\in\mathcal{B}(\mathbb{R})$, функції розподілу такі:

$$A)\qquad G(x)=\begin{cases}0, & x<0,\\ x, & 0\le x\le 1,\\ 1, & 1\le x,\end{cases}\qquad H(x)=\begin{cases}0, & x<0.25,\\ 2(x-0.25), & 0.25\le x\le 0.75,\\ 1, & 0.75\le x,\end{cases}$$

$$B)\qquad G(x)=\begin{cases}0, & x<0,\\ x, & 0\le x\le 1,\\ 1, & 1\le x,\end{cases}\qquad H(x)=\begin{cases}0, & x<1,\\ x-1, & 1\le x\le 2,\\ 1, & 2\le x.\end{cases}$$

З використанням пакету Mathematica 5 була змодельована суміш компонент з даними розподілами з функцією концентрації $w(t)$. По суміші була побудована оцінка $u_N^*(t)$ і обчислено $\min_{\sigma=\pm1}\|u-\sigma u_N^*\|_2$. Усереднені результати 30–50 експериментів для кожного набору параметрів наведені в таблиці. В першій колонці (під титулом N) йде довжина вибірки, по яким будувались оцінки, титули ($l=5,10,\dots$) відповідають кількості індикаторних функцій в базисі, по якому будується V_l (тобто $V_l=$ з.л.о.$\{g_k,k=1,\dots,l\}$, $g_k(x)=\sqrt{l}\,\mathbb{1}\{x\in(\frac{k-1}{l},\frac{k}{l}]\}$), буквами A, B помічено колонки з даними для відповідних прикладів. В першій стрічці під титулами $5,10,\dots$

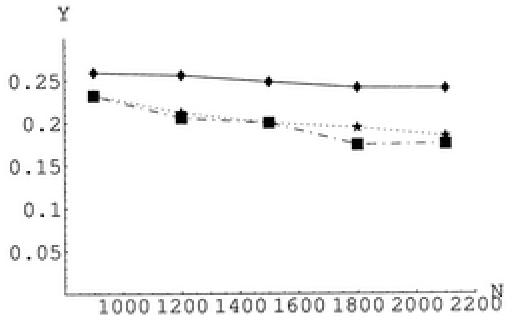

Рис. 2. Усереднена відстань $\min_{\sigma=\pm1} \|u - \sigma u_N^*\|_2$ для прикладу B

йдуть відстані $\|u - \mathrm{Pr}_{V_l} u\|_2$.

N	5		10		15		20	
	0.440		0.223		0.149		0.112	
	A	B	A	B	A	B	A	B
900	0.497	0.463	0.397	0.259	0.424	0.233	0.459	0.232
1200	0.477	0.459	0.345	0.256	0.367	0.212	0.371	0.206
1500	0.472	0.458	0.302	0.248	0.317	0.200	0.315	0.200
1800	0.464	0.456	0.302	0.241	0.301	0.194	0.294	0.174
2100	0.466	0.454	0.282	0.240	0.255	0.184	0.284	0.175

На графіках показані результати для обох прикладів для трьох значень l: 10 — ромбами, 15 — зірочками, 20 — квадратиками. Відстані відкладені по осі Y, величини вибірки — по N.

4. Висновки

Таким чином, ми розглянули питання про побудову проекційних оцінок форми функції концентрації з розкладом за гістограмним базисом для даних, що являють собою вибірку з суміші двох компонент з невідомими розподілами, концентрація яких змінюється протягом спостережень. Доведено консистентність оцінок і оцінено швидкість збіжності м.н. Показано, що швидкість збіжності для оцінок, побудованих за гістограмним базисом має той самий порядок, що і для оцінок, побудованих за тригонометричним базисом.

Результати імітаційного моделювання показали, що у випадку розподілів, які легко розрізняються (приклад B), величина похибки досить помірна вже при порівняно невеликих обсягах вибірки, зменшення похибки відбувається головним чином при збільшенні кількості базисних функцій. У випадку, коли розподіли розрізняються гірше (приклад A), похибка "стартує" з досить великих значень (майже 0.5 норми функції, що оцінюється), проте із зростанням обсягу вибірки помітно зменшується: ефект зменшення похибки при збільшенні кількості елементів гістограмного базису проявляється значно слабше, ніж в прикладі B.

Отримані величини похибок не виглядають малими, проте потрібно врахувати те, що дана оцінка використовує мінімальні припущення — припускається лише те, що функції розподілу є різними відносно деякої заданої міри і ліпшицевість функції концентрації.

Література

1. Р. Є. Майборода, *Асимптотично ефективна оцінка ймовірностей, побудована по спостереженням із суміші*, Теор. ймовірност. матем. статист. **59** (1998), 121–128.
2. Р. Є. Майборода, *Непараметрична статистика неоднорідних спостережень*, Дисертація на здобуття вченого ступеня доктора фізико-математичних наук, Київ, 1994.
3. Р. Є. Майборода, *Проекційні оцінки концентрацій сумішей що змінюються*, Теор. ймовірност. матем. статист. **46** (1992), 70–76.
4. Р. Є. Майборода, *Оцінка розподілів компонентів сумішей з концентраціями, що змінюються*, Укр. мат. ж. **48** (1996), № 4, 562–566.
5. Р. Є. Майборода, *Статистичний аналіз сумішей*, Навчальний посібник, Видавничо-поліграфічний центр "Київський університет", Київ, 2003.
6. К. Чуи, *Введение в вэйвлеты*, "Мир", Москва, 2001.
7. I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1996.
8. W. Härdle, *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, 1990.
9. W. Härdle, G. Kerkyacharian, D. Picard, A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*, Springer-Verlag, New York, 1998.
10. R. E. Maiboroda, *Estimation and classification by mixtures with time-dependent concentrations*, VI International Vilnius Conference on Probability Theory and Math. Statistics. Abstracts of Communications, т. 2, 1993, стор. 48.
11. O. V. Sugakova, *Asymptotics of a kernel estimate for distribution density constructed from observations of a mixture with varying concentration*, Theor. Probab. Math. Statist. **59** (1999), 161–171.
12. B. Vidakovic, *Statistical Modeling by Wavelets*, John Wiley & Sons, New York, 1999.

Кафедра теорії ймовірностей та математичної статистики, механіко-математичний факультет, Національний університет ім. Тараса Шевченка, проспект Глушкова, 6, Київ 03127, Україна

Адреса електронної пошти: `pokhid@ukr.net`

Надійшла 26/05/2004